REMARKS

This communication is in response to the Office Action dated December 19, 2003. Claims 1-12 are pending in the present Application. Claims 1-12 have been rejected. Claims 1-12 remain pending in the present Application.

The present invention is a method for operating a server cluster that includes N Server nodes that service client requests. Each client request is directed to one of a plurality of sites hosted on the server cluster. Each site is identified by a domain name and each server node is identified by an address on a network connecting the clients to the server nodes. In an embodiment, the computational resources required to service the requests to each of the sites over a first time period are measured and used to group the sites into N groups.

Each group is assigned to a corresponded one of the server nodes. The groups are chosen such that for each pair of groups, a difference in the sum of the measured computational resources is within a first predetermined error value. Configuration information defining a correspondence between each of the sites and one or more of the server nodes assigned to the groups containing that site is then provided to a router accessible from the network. The router provides the address of the server node that is to service the next client request for each site.

§112 Rejections

The Examiner states:

> **Claims 1-12 are rejected under 35 U.S.C. 112, second paragraph as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.**
>
> **Claim 1 recites the limitation "the difference in sum" in claim 1, page 25, line 13. There is insufficient antecedent basis for this limitation in the claim.**

Applicant respectfully asserts that Claim 1 has been amended to overcome the above-referenced informality. Specifically, Claim 1 has been amended to recite "...a difference in the sum...".

§103 Rejections

Claims 1-4 and 6-12

For ease of review, Applicant reproduces independent Claim 1 herein below:

1. A method for operating a server cluster comprising N server nodes to service client requests, each client request being directed to one of a plurality of sites hosted on said server cluster, each site being identified by a domain name and each server node being identified by an address on a network connecting said clients to said server nodes, said method comprising the steps of:

measuring the computational resources required to service said requests to each of said sites over a first time period;

grouping said sites into N groups, each group being assigned to a corresponding one of said server nodes such that for each pair of groups, a difference in the sum of said measured computational resources is within a first predetermined error value; and

providing configuration information to a router accessible from said network, said information defining a correspondence between each of said sites and one of said server nodes assigned to one of said groups containing that site, said router providing said address of said server node in response to a message specifying said domain name of said site.

The Examiner states:

> **Claims 1-4 and 6-12 are rejected under 35 U.S.C. 103(a) as being unpatentable over Yu, US Patent No. 6351775 and further in view of DeBettencourt US 6,279,001.**

Applicant respectfully disagrees. Here, the Examiner is attempting to combine the Yu reference with the DeBettencourt reference. When making a rejection under 35 U.S.C. § 103, a necessary condition is that the combination of the cited references ***must teach or suggest all claim limitations***. If the cited references do not teach or suggest every element of the claimed invention, then the cited references fail to render obvious the claimed invention, i.e. the claimed invention is distinguishable over the combination of the cited references. (Emphasis added.)

Claim 1 recites "...measuring the computational resources required to service said requests to each of said sites over a first time period... such that for each pair of groups, a difference in the sum of said measured computational resources is within a first predetermined error value ...". The Examiner admits that Yu does not disclose this limitation and purports that the DeBettencourt reference provides this limitation. The Examiner specifically points to col. 22, line 37-col. 23, line 33 of DeBettencourt reference which is reproduced below:

> The interceptor 120 chooses which web server 102 it will refer a request to based on a load metric ("LM") determined for each available web server 102. Each web server 102 is mapped to an interval between 0 and 1. The size of the interval associated with a web server 102 is proportional to the load metric for that web server 102. The interceptor 120 generates a random number between 0 and 1. The web server 102 mapped to the interval containing the chosen random number is selected as the web server 102 that will receive the request. In this way, there is a somewhat random distribution, yet there is a higher probability that the web servers 102 with the lightest load will be chosen.
>
> For example, and referring to FIG. 12A, if there are six web servers A, B, C, D, E and F, each of the six web servers A-F will be assigned to an interval between 0 and 1. The width of the interval will be proportional to the weighted load metric for that web server. In this example, the six web servers have the load metrics LM_A =1500, LM_B =2250, LM_C =3250, LM_D =2000, LM_E =1000, and LM_F =1000. The load metrics total 10,000, so to normalize the intervals to a range between 0 and 1, each load metric is divided by 10,000. This produces the following interval widths ("W") for each web server: W_A =0.150, W_B =0.225, W_C =0.325, W_D =0.2, W_E =0.1, and W_F =0.1. Each web server is assigned an interval that is of the appropriate width in the range between 0 and 1. In this example, web server A is assigned the interval 0-0.150, web server B is assigned the interval 0.15-0.375, web server C is assigned the interval 0.375-0.6, web server D is assigned the interval 0.601-0.800, web server E is assigned the interval 0.801-0.9, and web server F is assigned the interval 0.901-1.0. Referring

> to FIG. 12B, the mapping of these intervals to the range 0 to 1 shows that the intervals cover the range 0 to 1. As is apparent from the figure, web server C, which in this example has the largest weighted load value, LM_C =3250, indicating that this web server can process requests most quickly, has the largest interval, W_C =0.325. Web server C has a high probability of receiving new requests.
>
> Having distributed the web servers on the interval, the interceptor 120 generates a random number between 0 and 1. In this example, the interceptor 120 generates the random number 0.517. The interceptor 120 sends the request to the web server 102 that has the interval that contains the number 0.517. In this example, the number 0.517 falls into the range 0.376-0.6, and so the request is referred to web server C.
>
> The Load Metric
>
> In one embodiment, the load metric for each web server is determined by a static, default capacity value ("C"). The default capacity value can be assigned by the system operator to each web server 102 in the web service system 90. In one embodiment, the system operator can assign a value ranging from 1 to 10 to each web server 102, which is a relative evaluation of the load capacity of that web server 102. For example, the web server 102 with the greatest capacity, possibly with a relatively large number of processors running at the relatively high clock speed, can be assigned a capacity of 10. A relatively slow web server 102 with only one processor can be assigned a capacity of 1.
>
> In another embodiment, the load metric for each web server 102 is determined by a dynamic load value generated by the manager 110. The manager 110 periodically sends an updated load value for each web server 102 to the interceptor 120. The dynamic load value reflects the current capacity of each web server 102 based on one or more metrics that provide real-time evaluation of web server performance.

Applicant assert that the above-disclosed portion of the Debettencourt reference does not teach or suggest the limitation "... such that for each pair of groups, a difference in the sum of said measured computational resources is within a first predetermined error value ..." as recited in Claim 1 of the present invention.

Specifically, the above-disclosed portion of Debettencourt discloses two embodiments for determining a load metric for each available web server. In a first embodiment, the load metric for each web server is determined by a static, default capacity value. In the second embodiment, the load metric for each web server is determined by a dynamic load value generated by the manager wherein the manager periodically sends an updated load value for each web server to the interceptor. ***Neither embodiment disclosed in the Debettencourt reference teaches or suggests that a difference in the sum of measured computational resources is within a first predetermined error value as recited in Claim 1 of the present invention.***

Consequently, since neither embodiment disclosed in the Debettencourt reference teaches or suggests that a difference in the sum of measured computational resources is within a first predetermined error value as recited in Claim 1 of the present invention, the Examiner's proposed Yu-Debettencourt combination does not teach or suggest every limitation recited in claim 1 of the present invention. Accordingly, since the Examiner's proposed Yu-Debettencourt combination does not teach or suggest every limitation recited in Claim 1 of the present invention, Claim 1 is allowable over the Examiner's proposed combination of references.

Since Claims 2-4 and 6-12 are dependent on Claim 1, the above-articulated argument with regard to Claim 1 applies with equal force to claims 2-4 and 6-12. Accordingly, Claim 2-4 and 6-12 should be allowed over this reference.

Claim 5

The Examiner states:

> **Claims 18-23 and 27-31 are rejected under 35 U.S.C. 103(a) as being unpatentable over Gabriel et al. (US 4754185) and Hayashi et al. (Hayashi) (JP 06113563).**

Since Claim 5 is dependent on Claim 1, the above-articulated argument with regard to Claim 1 applies with equal force to Claim 5. Applicant further asserts that the Desai reference fails to correct the outlined deficiencies of the Yu-DeBettencourt et al.

combination of references. Accordingly, since the Desai reference fails to correct the outlined deficiencies of the Yu-DeBettencourt et al. combination of references, Claim 5 should be allowed over the Examiner's proposed combination of references.

Applicant believes that this application is in condition for allowance. Accordingly, Applicant respectfully requests reconsideration, allowance and passage to issue of the claims as now presented. Should any unresolved issues remain, Examiner is invited to call Applicant's attorney at the telephone number indicated below.

Respectfully submitted,

Wendell J. Jones
Attorney for Applicant
Reg. No. 45,961
(408) 938-0980